Review

Statistics review 2: Samples and populations

Elise Whitley* and Jonathan Ball†

*Lecturer in Medical Statistics, University of Bristol, UK
†Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 7 February 2002 *Critical Care* 2002, **6**:143-148

Abstract

The previous review in this series introduced the notion of data description and outlined some of the more common summary measures used to describe a dataset. However, a dataset is typically only of interest for the information it provides regarding the population from which it was drawn. The present review focuses on estimation of population values from a sample.

Keywords confidence interval, normal distribution, reference range, standard error

In medical (and other) research there is generally some population that is ultimately of interest to the investigator (e.g. intensive care unit [ICU] patients, patients with acute respiratory distress syndrome, or patients who receive renal replacement therapy). It is seldom possible to obtain information from every individual in the population, however, and attention is more commonly restricted to a sample drawn from it. The question of how best to obtain such a sample is a subject worthy of discussion in its own right and is not covered here. Nevertheless, it is essential that any sample is as representative as possible of the population from which it is drawn, and the best means of obtaining such a sample is generally through random sampling. (For more details see Bland [1].)

Once a (representative) sample has been obtained it is important to describe the data using the methods described in Statistics review 1. However, interest is rarely focused on the sample itself, but more often on the information that the sample can provide regarding the population of interest.

The Normal distribution

Quantitative clinical data follow a wide range of distributions. By far the most common of these is symmetrical and unimodal, with a single peak in the middle and equal tails at either side. This distinctive bell-shaped distribution is known as 'Normal' or 'Gaussian'. Note that Normal in this context (written with an upper case 'N') has no implications in terms of clinical normality, and is used purely to describe the shape of the distribution.

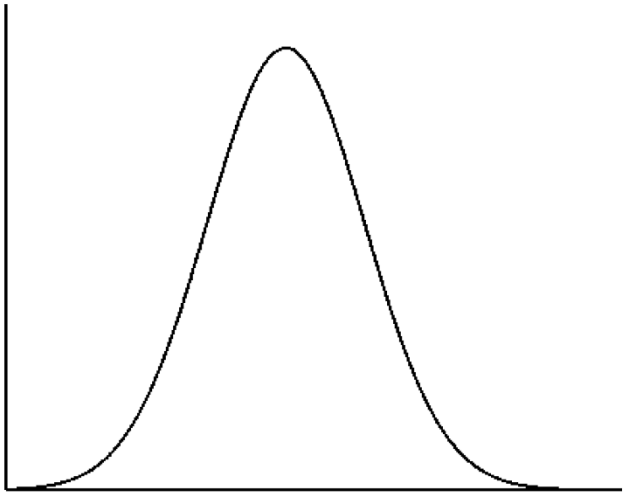
Strictly speaking, the theoretical Normal distribution is continuous, as shown in Fig. 1. However, data such as those shown in Fig. 2, which presents admission haemoglobin concentrations from intensive care patients, often provide an excellent approximation in practice.

There are many other theoretical distributions that may be encountered in medical data, for example Binary or Poisson [2], but the Normal distribution is the most common. It is additionally important because it has many useful properties and is central to many statistical techniques. In fact, it is not uncommon for other distributions to tend toward the Normal distribution as the sample size increases, meaning that it is often possible to use a Normal approximation. This is the case with both the Binary and Poisson distributions.

One of the most important features of the Normal distribution is that it is entirely defined by two quantities: its mean and its standard deviation (SD). The mean determines where the peak occurs and the SD determines the shape of the curve. For example, Fig. 3 shows two Normal curves. Both have the same mean and therefore have their peak at the same value. However, one curve has a large SD, reflecting a large amount of deviation from the mean, which is reflected in its short, wide shape. The other has a small SD, indicating that individual values generally lie close to the mean, and this is reflected in the tall, narrow distribution.

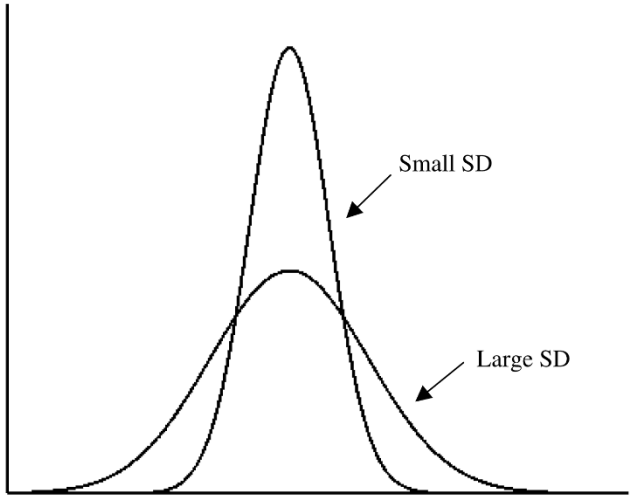
ICU=intensive care unit; SD=standard deviation; SE=standard error.

Figure 1



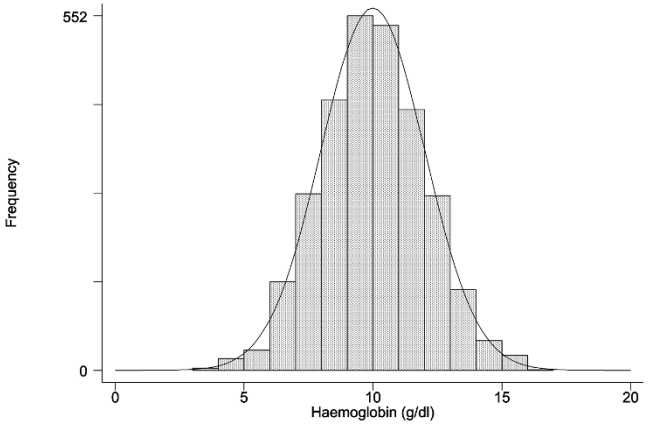
The Normal distribution.

Figure 3



Normal curves with small and large standard deviations (SDs).

Figure 2



Admission haemoglobin concentrations from 2849 intensive care patients.

It is possible to write down the equation for a Normal curve and, from this, to calculate the area underneath that falls between any two values. Because the Normal curve is defined entirely by its mean and SD, the following rules (represented by parts a–c of Fig. 4) will always apply regardless of the specific values of these quantities: (a) 68.3% of the distribution falls within 1 SD of the mean (i.e. between mean – SD and mean + SD); (b) 95.4% of the distribution falls between mean – 2 SD and mean + 2 SD; (c) 99.7% of the distribution falls between mean – 3 SD and mean + 3 SD; and so on.

The proportion of the Normal curve that falls between other ranges (not necessarily symmetrical, as here) and, alternatively, the range that contains a particular proportion of the Normal curve can both be calculated from tabulated values [3]. However, one proportion and range of particular interest is as follows (represented by part d of Fig. 4); 95% of the distribution falls between mean – 1.96 SD and mean + 1.96 SD.

The standard deviation and reference range

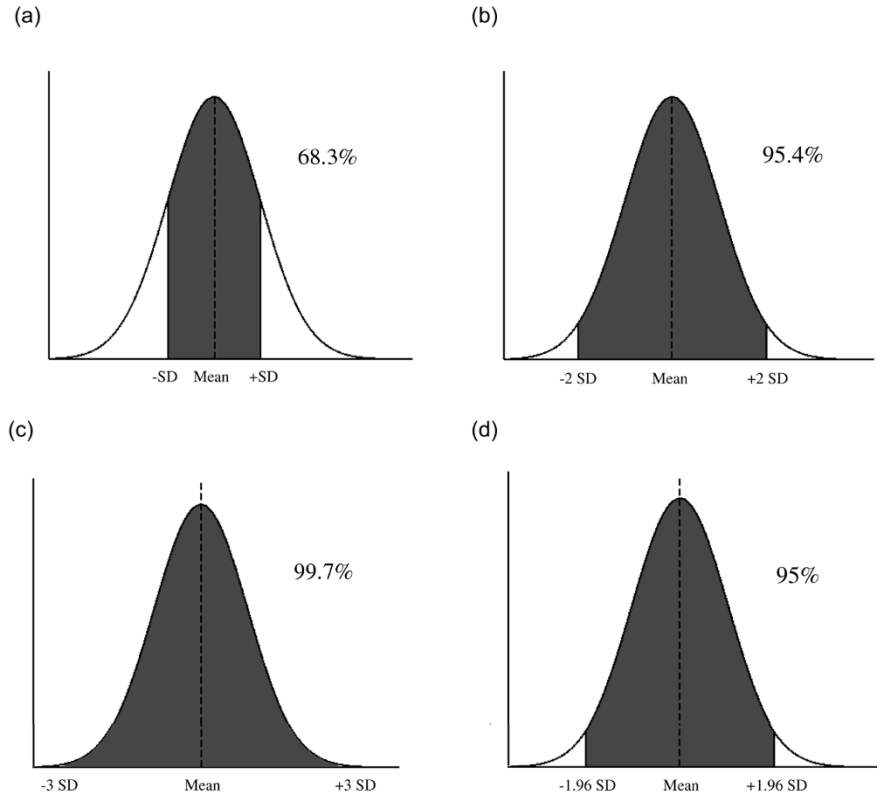
The properties of the Normal distribution described above lead to another useful measure of variability in a dataset. Rather than using the SD in isolation, the 95% reference range can be calculated as (mean – 1.96 SD) to (mean + 1.96 SD), provided that the data are (approximately) Normally distributed. This range will contain approximately 95% of the data. It is also possible to define a 90% reference range, a 99% reference range and so on in the same way, but conventionally the 95% reference range is the most commonly used.

For example, consider admission haemoglobin concentrations from a sample of 48 intensive care patients (see Statistics review 1 for details). The mean and SD haemoglobin concentration are 9.9 g/dl and 2.0 g/dl, respectively. The 95% reference range for haemoglobin concentration in these patients is therefore:

$(9.9 - [1.96 \times 2.0])$ to $(9.9 + [1.96 \times 2.0]) = 5.98$ to 13.82 g/dl.

Thus, approximately 95% of all haemoglobin measurements in this dataset should lie between 5.98 and 13.82 g/dl. Comparing this with the measurements recorded in Table 1 of Statistics review 1, there are three observations outside this range. In other words, 94% (45/48) of all observations are within the reference range, as expected.

Figure 4



Areas under the Normal curve. Because the Normal distribution is defined entirely by its mean and standard deviation (SD), the following rules apply: (a) 68.3% of the distribution falls within 1 SD of the mean (i.e. between mean – SD and mean + SD); (b) 95.4% of the distribution falls between mean – 2 SD and mean + 2 SD; (c) 99.7% of the distribution falls between mean – 3 SD and mean + 3 SD; and (d) 95% of the distribution falls between mean – 1.96 SD and mean + 1.96 SD.

Now consider the data shown in Fig. 5. These are blood lactate measurements taken from 99 intensive care patients on admission to the ICU. The mean and SD of these measurements are 2.74 mmol/l and 2.60 mmol/l, respectively, corresponding to a 95% reference range of −2.36 to +7.84 mmol/l. Clearly this lower limit is impossible because lactate concentration must be greater than 0, and this arises because the data are not Normally distributed. Calculating reference ranges and other statistical quantities without first checking the distribution of the data is a common mistake and can lead to extremely misleading results and erroneous conclusions. In this case the error was obvious, but this will not always be the case. It is therefore essential that any assumptions underlying statistical calculations are carefully checked before proceeding. In the current example a simple transformation (e.g. logarithmic) may make the data approximately Normal, in which case a reference range could legitimately be calculated before transforming back to the original scale (see Statistics review 1 for details).

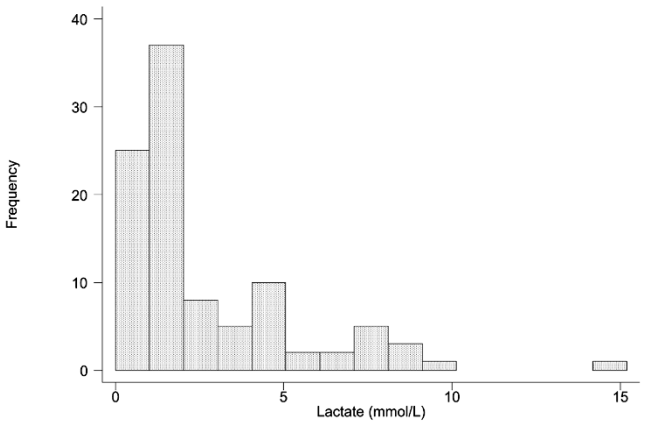
Two quantities that are related to the SD and reference range are the standard error (SE) and confidence interval. These quantities have some similarities but they measure very different things and it is important that they should not be confused.

From sample to population

As mentioned above, a sample is generally collected and calculations performed on it in order to draw inferences regarding the population from which it was drawn. However, this sample is only one of a large number of possible samples that might have been drawn. All of these samples will differ in terms of the individuals and observations that they contain, and so an estimate of a population value from a single sample will not necessarily be representative of the population. It is therefore important to measure the variability that is inherent in the sample estimate. For simplicity, the remainder of the present review concentrates specifically on estimation of a population mean.

Consider all possible samples of fixed size (n) drawn from a population. Each of these samples has its own mean and these means will vary between samples. Because of this variation, the sample means will have a distribution of their own. In fact, if the samples are sufficiently large (greater than

Figure 5



Lactate concentrations in 99 intensive care patients.

approximately 30 in practice) then this distribution of sample means is known to be Normal, regardless of the underlying distribution of the population. This is a very powerful result and is a consequence of what is known as the Central Limit Theorem. Because of this it is possible to calculate the mean and SD of the sample means.

The mean of all the sample means is equal to the population mean (because every possible sample will contain every individual the same number of times). Just as the SD in a sample measures the deviation of individual values from the sample mean, the SD of the sample means measures the deviation of individual sample means from the population mean. In other words it measures the variability in the sample means. In order to distinguish it from the sample SD, it is known as the standard error (SE). Like the SD, a large SE indicates that there is much variation in the sample means and that many lie a long way from the population mean. Similarly, a small SE indicates little variation between the sample means. The size of the SE depends on the variation between individuals in the population and on the sample size, and is calculated as follows:

$$SE = \sigma/\sqrt{n} \quad (1)$$

where σ is the SD of the population and n is the sample size. In practice, σ is unknown but the sample SD will generally provide a good estimate and so the SE is estimated by the following equation:

$$SE = \text{Sample SD}/\sqrt{n} \quad (2)$$

It can be seen from this that the SE will always be considerably smaller than the SD in a sample. This is because there is less variability between the sample means than between individual values. For example, an individual admission haemoglobin level of 8 g/dl is not uncommon, but to obtain a sample of 100 patients with a mean haemoglobin level of 8 g/dl would require the majority to have scores well below average, and this is unlikely to occur in practice if the sample is truly representative of the ICU patient population.

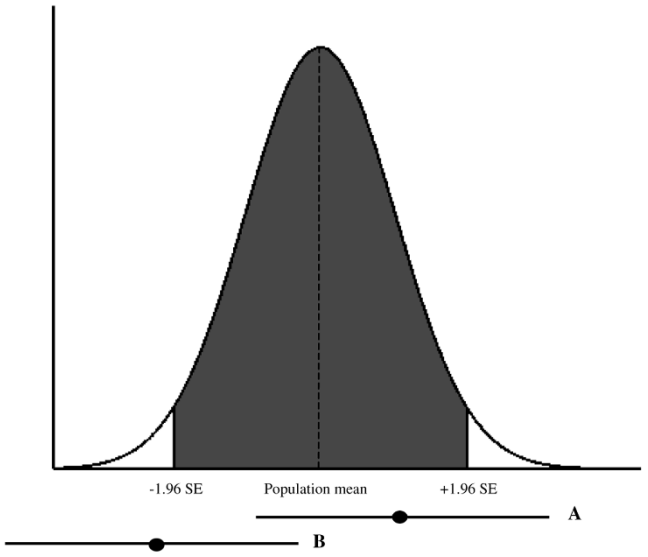
It is also clear that larger sample sizes lead to smaller standard errors (because the denominator, $\sqrt{n}$, is larger). In other words, large sample sizes produce more precise estimates of the population value in question. This is an important point to bear in mind when deciding on the size of sample required for a particular study, and will be covered in greater detail in a subsequent review on sample size calculations.

The standard error and confidence interval

Because sample means are Normally distributed, it should be possible to use the same theory as for the reference range to calculate a range of values in which 95% of sample means lie. In practice, the population mean (the mean of all sample means) is unknown but there is an extremely useful quantity, known as the 95% confidence interval, which can be obtained in the same way. The 95% confidence interval is invaluable in estimation because it provides a range of values within which the true population mean is likely to lie. The 95% confidence interval is calculated from a single sample using the mean and SE (derived from the SD, as described above). It is defined as follows: (sample mean – 1.96 SE) to (sample mean + 1.96 SE).

To appreciate the value of the 95% confidence interval, consider Fig. 6. This shows the (hypothetical) distribution of sample means centred around the population mean. Because the SE is the SD of the distribution of all sample means, approximately 95% of all sample means will lie within 1.96 SEs of the (unknown) population mean, as indicated by the shaded area. A 95% confidence interval calculated from a sample with a mean that lies within this shaded area (e.g. confidence interval A in Fig. 6) will contain the true population mean. Conversely, a 95% confidence interval based on a sample with a mean outside this area (e.g. confidence interval B in Fig. 6) will not include the population mean. In practice it is impossible to know whether a sample falls into the first or second category; however, because 95% of all sample means fall into the shaded area, a confidence interval that is based on a single sample is likely to contain the true population mean 95% of the time. In other words, given a 95% confidence interval based on a single sample, the investigator can be 95% confident that the true population mean (i.e. the real measurement of interest) lies somewhere within that range. Equally important is that 5% of such intervals will not contain the true population value. However, the choice of 95% is purely arbitrary, and using a 99% confidence interval (calculated as mean ± 2.56 SE) instead will make it more likely that the true value is contained within the range. However, the cost of this change is that the range will be wider and therefore less precise.

Figure 6



The distribution of sample means. The shaded area represents the range of values in which 95% of sample means lie. Confidence interval A is calculated from a sample with a mean that lies within this shaded area, and contains the true population mean. Confidence interval B, however, is calculated from a sample with a mean that falls outside the shaded area, and does not contain the population mean. SE=standard error.

As an example, consider the sample of 48 intensive care patients whose admission haemoglobin concentrations are described above. The mean and SD of that dataset are 9.9 g/dl and 2.0 g/dl, respectively, which corresponds to a 95% reference range of 5.98 to 13.82 g/dl. Calculation of the 95% confidence interval relies on the SE, which in this case is $2.0/\sqrt{48} = 0.29$. The 95% confidence interval is then:

$(9.9 - [1.96 \times 0.29])$ to $(9.9 + [1.96 \times 0.29]) = 9.33$ to 10.47 g/dl

So, given this sample, it is likely that the population mean haemoglobin concentration is between 9.33 and 10.47 g/dl. Note that this range is substantially narrower than the corresponding 95% reference range (i.e. 5.98 to 13.82 g/dl; see above). If the sample were based on 480 patients rather than just 48, then the SE would be considerably smaller ($SE = 2.0/\sqrt{480} = 0.09$) and the 95% confidence interval (9.72 to 10.08 g/dl) would be correspondingly narrower.

Of course a confidence interval can only be interpreted in the context of the population from which the sample was drawn. For example, a confidence interval for the admission haemoglobin concentrations of a representative sample of postoperative cardiac surgical intensive care patients provides a range of values in which the population mean admission haemoglobin concentration is likely to lie, in postoperative cardiac surgical intensive care patients. It does not provide information on the likely range of admission haemoglobin concentrations in medical intensive care patients.

Confidence intervals for smaller samples

The calculation of a 95% confidence interval, as described above, relies on two assumptions: that the distribution of sample means is approximately Normal and that the population SD can be approximated by the sample SD. These assumptions, particularly the first, will generally be valid if the sample is sufficiently large. There may be occasions when these assumptions break down, however, and there are alternative methods that can be used in these circumstances. If the population distribution is extremely non-Normal and the sample size is very small then it may be necessary to use non-parametric methods. (These will be discussed in a subsequent review.) However, in most situations the problem can be dealt with using the t-distribution in place of the Normal distribution.

The t-distribution is similar in shape to the Normal distribution, being symmetrical and unimodal, but is generally more spread out with longer tails. The exact shape depends on a quantity known as the 'degrees of freedom', which in this context is equal to the sample size minus 1. The t distribution for a sample size of 5 (degrees of freedom = 4) is shown in comparison to the Normal distribution in Fig. 7, in which the longer tails of the t-distribution are clearly shown. However, the t-distribution tends toward the Normal distribution (i.e. it becomes less spread out) as the degrees of freedom/sample size increase. Fig. 8 shows the t-distribution corresponding to a sample size of 20 (degrees of freedom = 19), and it can be seen that it is already very similar to the corresponding Normal curve.

Calculating a confidence interval using the t-distribution is very similar to calculating it using the Normal distribution, as described above. In the case of the Normal distribution, the calculation is based on the fact that 95% of sample means fall within 1.96 SEs of the population mean. The longer tails of the t-distribution mean that it is necessary to go slightly further away from the mean to pick up 95% of all sample means. However, the calculation is similar, with only the figure of 1.96 changing. The alternative multiplication factor depends on the degrees of freedom of the t-distribution in question, and some typical values are presented in Table 1.

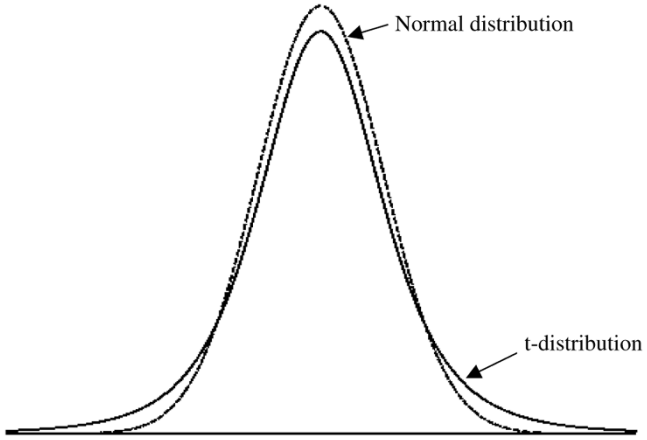
As an example, consider the admission haemoglobin concentrations described above. The mean and SD are 9.9 g/dl and 2.0 g/dl, respectively. If the sample were based on 10 patients rather than 48, it would be more appropriate to use the t-distribution to calculate a 95% confidence interval. In this case the 95% confidence interval is given by the following: mean ± 2.26 SE. The SE based on a sample size of 10 is 0.63, and so the 95% confidence interval is 8.47 to 11.33 g/dl.

Table 1

Multiplication factors for confidence intervals based on the t-distribution

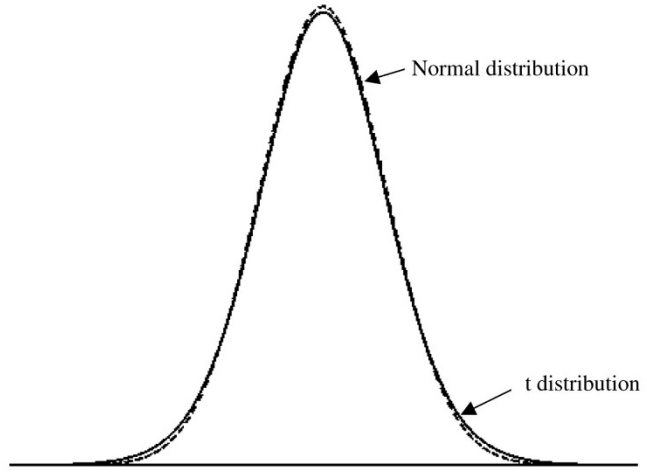
Sample size	10	20	30	40	50	200
Multiplication factor	2.26	2.09	2.05	2.02	2.01	1.97

Figure 7



The Normal and t (with 4 degrees of freedom) distributions.

Figure 8



The Normal and t (with 19 degrees of freedom) distributions.

Note that as the sample sizes increase the multiplication factors shown in Table 1 decrease toward 1.96 (the multiplication factor for an infinite sample size is 1.96). The larger multiplication factors for smaller samples result in a wider confidence interval, and this reflects the uncertainty in the estimate of the population SD by the sample SD. The use of the t-distribution is known to be extremely robust and will therefore provide a valid confidence interval unless the population distribution is severely non-Normal.

Standard deviation or standard error?

There is often a great deal of confusion between SDs and SEs (and, equivalently, between reference ranges and confidence intervals). The SD (and reference range) describes the amount of variability between individuals within a single sample. The SE (and confidence interval) measures the precision with which a population value (i.e. mean) is estimated by a single sample. The question of which measure to use is well summed up by Campbell and Machin [4] in the following mnemonic: "If the purpose is Descriptive use standard Deviation; if the purpose is Estimation use standard Error."

Confidence intervals are an extremely useful part of any statistical analysis, and are referred to extensively in the remaining reviews in this series. The present review concentrates on calculation of a confidence interval for a single mean. However, the results presented here apply equally to population proportions, rates, differences, ratios and so on. For details on how to calculate appropriate SEs and confidence intervals, refer to Kirkwood [2] and Altman [3].

Key messages

The SD and 95% reference range describe variability within a sample. These quantities are best used when the objective is description.

The SE and 95% confidence interval describe variability between samples, and therefore provide a measure of the precision of a population value estimated from a single sample. In other words, a 95% confidence interval provides a range of values within which the true population value of interest is likely to lie. These quantities are best used when the objective is estimation.

Competing interests

None declared.

References

1. Bland M: *An Introduction to Medical Statistics*. 3rd ed. Oxford, UK: Oxford University Press; 2001.
2. Kirkwood BR: *Essentials of Medical Statistics*. London, UK: Blackwell Science Ltd; 1988.
3. Altman DG: *Practical Statistics for Medical Research*. London, UK: Chapman & Hall; 1991.
4. Campbell MJ, Machin D: *Medical Statistics: a Commonsense Approach.* 2nd ed. Chichester, UK: John Wiley & Sons Ltd; 1993.